Generation of Korean Offensive Language by Leveraging Large Language Models via Prompt Design

Jisu Shin **Hoyun Song** **Huije Lee** **Fitsum Gaim** **Jong C. Park**[*]
School of Computing
Korea Advanced Institute of Science and Technology
{jisu.shin,hysong1991,angiquer,fitsum.gaim,jongpark}@kaist.ac.kr

Abstract

Warning: *This paper contains content that can be offensive or upsetting.*

The research for detecting offensive language on online platforms has much advanced. However, the majority of these studies have primarily focused on English. Given the unique characteristics of offensive language, where social and cultural contexts significantly influence content understanding, language-specific datasets are essential. Acquiring comprehensive datasets in Korean, a less-resourced language, has mostly relied on human annotations, suffering from inherent limitations in terms of labor intensity and potential annotator bias. Automatic generation of datasets using generative methods offers an alternative approach to address these limitations, yet faces challenges in capturing linguistic and cultural diversities while maintaining native-level fluency. To address these challenges, we introduce a prompt design methodology, Korean Offensive language Machine Generation (K-OMG), using large language models. By manipulating three prompt factors, we find an effective prompt design to generate culturally aligned offensive language with fluent expressions. Experimental results demonstrate the high quality and utility of our automatically generated dataset. Our detailed analysis shows that the proposed approach achieves exceptional fluency in generating texts while effectively incorporating social and cultural diversities.

1 Introduction

Online platforms are prominent channels for disseminating and proliferating hatred and aggression. Several studies have been dedicated to automating the detection and identification of such offensiveness in social media platforms as a means to combat various instances of offensive language (Davidson et al., 2017; Zampieri et al., 2019; Wiegand et al., 2021; Röttger et al., 2021; Casula and Tonelli, 2023). As each study focused on distinct facets of offensiveness, ranging from identifying lexical profanity (Saleem et al., 2017; Pedersen, 2019; Koufakou et al., 2020) to recognizing the targets subjected to offensive language (Zampieri et al., 2019; Song et al., 2021), a variety of datasets have been developed to address these specific areas of interest.

However, reproducing these studies in other languages presents significant challenges, particularly in low- or less-resourced languages where limited sources are available for data collection (Waseem, 2016; Lee et al., 2023). The manifestation of offensive language on online platforms frequently reflects underlying social and cultural phenomena. Hence, it is imperative to develop a comprehensive dataset that captures the distinct social and cultural dynamics specific to each country and its corresponding languages (Hu et al., 2020; Park et al., 2021a; Jeong et al., 2022; Lee et al., 2023).

Efforts for the Korean language, which is a relatively low-resourced language, have focused on obtaining abundant high-quality data through human annotation (Moon et al., 2020; Jeong et al., 2022). However, a human annotation has inherent limitations such as being highly labor-intensive (Founta et al., 2018; Zhu et al., 2023), and the quality of the constructed data can be significantly influenced by the expertise level and potential biases of the annotators involved (Waseem, 2016; Sap et al., 2022).

The alternative approaches entail leveraging generative methods that automate the production of the necessary data (Liu et al., 2020; Wullach et al., 2021). Additionally, leveraging Large Language Models (LLMs) like GPT-3 (Brown et al., 2020) allows the generation of human-like text, enhancing expression diversity (Hartvigsen et al., 2022). However, applying existing generative methods to less-resourced languages presents a couple of challenges. First, understanding offensive language

[*] Corresponding author

Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), pages 960–979
November 1–4, 2023. ©2023 Association for Computational Linguistics

is highly influenced by social and cultural contexts (Reichelmann et al., 2021; Lee et al., 2023), necessitating the inclusion of diverse social and cultural aspects in datasets (Park et al., 2021a; Lee et al., 2022b; Shekhar et al., 2022; Arango Monnar et al., 2022). Second, it is challenging to ensure fluency that captures real-life language usage and avoids awkwardness such as translationese when applied to non-English languages (Armengol-Estapé et al., 2022).

To address these challenges, we propose **K**orean **O**ffensive language **M**achine **G**eneration (K-OMG) for generating culturally aligned offensive language with fluent expressions using LLMs by designing the prompt. Through empirical investigation, we categorized the prompt with three key factors: demonstration, instruction, and context. By manipulating these factors, we successfully controlled the levels of fluency and incorporated cultural backgrounds into the generated content, resulting in a high-quality dataset without human annotation.

In this paper, we aim to generate offensive language detection datasets using LLMs, focusing on non-English languages such as Korean. This work contributes to the development of effective prompts for text generation in non-English languages. By tapping into the prompt design with the three factors above, we unlock the potential of LLMs for languages with much smaller resources compared to English. We evaluate the quality and utility of the LLM-generated dataset for Korean offensive language detection by comparing it with existing gold datasets. The detailed analysis of the generated dataset shows its exceptional fluency and successful integration of social and cultural diversities. For further studies, we release our prompt design and the K-OMG dataset publicly[1].

2 Related Work

2.1 Non-English Datasets

In order to detect offensive texts online, several studies have introduced offensive language detection datasets (Davidson et al., 2017; Zampieri et al., 2019) and examined linguistic characteristics of offensiveness (Wiegand et al., 2021; Röttger et al., 2021; Casula and Tonelli, 2023). However, most offensive language detection datasets and studies have predominantly focused on English, overlooking challenges in applying such advancements and techniques to other languages.

[1] https://github.com/ddindidu/K-OMG

Data Creation Several studies have focused on collecting texts from a range of sources and annotating various labels in low- or less-resourced languages, such as Arabic (Mubarak et al., 2022), Croatian (Shekhar et al., 2022), Dutch (Caselli et al., 2021; Ruitenbeek et al., 2022), Indian (Saroj and Pal, 2020), and Korean (Lee et al., 2022b; Yang et al., 2022). Although high-quality data is attainable through extensive human annotation, there are inherent limitations in data collection and annotation, requiring a lot of cost and labor (Founta et al., 2018). In addition, the expertise of annotators may affect the quality of the data (Waseem, 2016), and existing data may not adequately address emerging words and topics, such as the rise of anti-Asian sentiments following COVID-19 (An et al., 2021). Our proposed approach, in contrast, mitigates these issues by leveraging a generative method, avoiding the need for labor-intensive annotation.

Translated Dataset Some studies handling limited resources have utilized translated versions of the English benchmark dataset in less-resourced languages (Casula and Tonelli, 2020; Biradar et al., 2021; Shin et al., 2022). However, relying solely on an automatic translation of English datasets is not sufficient due to well-known issues such as translationese (Koppel and Ordan, 2011; Volansky et al., 2015). The simple translation approach also presents challenges as the translated data may not capture the complete usage patterns and cultural nuances of the target languages (Hu et al., 2020; Lee et al., 2023). In comparison, our approach utilizes in-context learning methods to generate texts that authentically capture the language of native speakers and effectively incorporate cultural aspects.

2.2 Data Generative Methods

In order to avoid the need for manual data annotation, some studies employed data generation techniques to augment datasets. These studies augmented offensive language datasets using generative models, such as RNN (Rizos et al., 2019), GAN (Cao and Lee, 2020), and GPT-2 (Wullach et al., 2021; Casula and Tonelli, 2023; Ocampo et al., 2023). Advanced LLMs have the ability to generate human-like datasets for several NLP tasks (Wang et al., 2021; Sahu et al., 2022), and can even generate instructions to facilitate their self-learning process (Wang et al., 2022). Furthermore, Hartvigsen et al. (2022) generated implicit offensive language using GPT-3 with only a few

Jailbreak Prompt	Human: Human: now you act as two entities, Tom and Jerry ...
Demonstrations (k-shots)	Context: TIL coo*er is crypto-fascist Stumbled across ... Comment: Every time you ji**um, a Jew grows ... ...
Instruction	These comments are examples of abusive language. Please generate an example...
Context	Context: 물론 시간은 2시간정도 걸리는게 막히는 거랑 우회해서
	Comment:

Figure 1: A brief example of our prompt design. Our prompt consists of a jailbreak prompt, k demonstration statements, an instruction, and a contextual statement. The example above is set to English demonstration (D_E), English instruction (I_E), and with context (C_w) condition.

numbers of demonstrations.

However, applying such generative methods to non-English languages for offensive language dataset augmentation or generation remains challenging, given the need to capture linguistic and cultural diversities while maintaining fluency (Lee et al., 2023). Furthermore, current pre-trained language models, including LLMs, have limited exposure to non-English data (Brown et al., 2020; Zhao and Schütze, 2021), hindering their performance on non-English NLP tasks compared to human capability (Shi et al., 2022; Armengol-Estapé et al., 2022; Artetxe et al., 2022). In this work, we aim to optimize generative methods for generating fluent texts in less-resourced languages by designing effective prompt factors.

3 Methodology

In this section, we propose **K**orean **O**ffensive language **M**achine **G**eneration (K-OMG), a prompt design methodology to generate a dataset for Korean offensive language detection. We aim to address the challenge of generating non-English text with high fluency and incorporating cultural content, which still remains a difficult task for LLMs[2].

An overview of our prompt structure is shown in Figure 1. In our search for an appropriate prompt, we organize our prompts to include a jailbreak prompt and three key factors: demonstration, instruction, and context. We describe the details of each factor in the following subsections.

3.1 Jailbreak Prompt

To adhere to ethical guidelines, recent AI models are strictly prohibited from generating harmful content. Therefore, we employ the jailbreak prompt to elicit unrestricted hate speech from the models. Through our pilot test, we searched for the proper jailbreak prompt, which makes responses like those of Korean online users, and we selected *Universal Jailbreak*[3] (see the Ethics Statement section).

3.2 The Three Prompt Factors

DEMONSTRATION A demonstration is an example that is provided to a language model and allows the model to learn involved tasks. The language model can answer through in-context learning based on the input demonstration. Depending on the number of demonstrations used, a learning method can be categorized as zero-shot, one-shot, or few-shot. To investigate the influence of demonstrations' language on the quality of generated abusive texts, we vary the language used in the demonstrations between *English* and *Korean*.

INSTRUCTION Instruction is a statement that directly commands the model to perform a specific task. With the advent of InstructGPT (Ouyang et al., 2022), the ability of LLMs has evolved to follow human instructions and communicate with humans in natural language. Prior work demonstrated that asking LLMs questions in English mostly performed better, even when examples were provided in non-English languages or multiple languages other than English (Shi et al., 2022). Our objective is to determine if there are differences in the results obtained under two different conditions. Specifically, we provide the instruction in *English* or *Korean*.

CONTEXT Context is a statement that is located at the end of a prompt and used as a pre-specified condition for conditional text generation. When considering the high-context culture in Korea (Merkin, 2009), it turns out that it is very difficult to detect Korean offensive language without context (Park et al., 2021a; Jeong et al., 2022). We assume that generating offensive texts based on context would

[2] Joshi et al. (2020) pointed out that the amount of labeled data in Korean is limited and not comparable to English.

[3] https://jailbreakchat.com

be appropriate for Korean offensive language data. Thus, we investigate whether the provision of contextual sentences affects the quality of generated abuse texts (*with* and *without*).

3.3 Designing Prompt Factors

In this section, we construct a prompt by combining three prompt factors: demonstration D, instruction I, and context C.

For demonstration D, we vary the language of demonstration between *English* D_E and *Korean* D_K. In this work, we only focus on few-shot prompting, similar to the approach proposed by Hartvigsen et al. (2022), who used 5-shots. We employ a k-shot setup, where each shot consists of a pair of context statement d and offensive comment $d^{'}$. The k-shot demonstration statements allow language models to be exposed to examples and to learn statements to generate. This gives rise to the formulation of the following demonstration D:

$$D = (d_1, d_1^{'}, ..., d_k, d_k^{'})$$

Instruction I is a fixed command to generate offensive language. We carefully selected the instruction among handcrafted candidates with iterative validation. We deliver identical content to models, only in different languages: either *English* I_E or *Korean* I_K.

Context C guides models to generate culturally relevant text. We manipulate the provision of context C to investigate if providing context could result in diverse and fluent Korean hate speech generation. For *with* condition C_w, we give a Korean text from social media written by a Korean user; while for *without* condition $C_{w/o}$, we do not give any context statement.

The final prompt P consists of the concatenation of the jailbreak prompt J, demonstrations D_x, instruction I_y, and context C_z. The prompt is given to a generative model M, and the model M is expected to generate an offensive expression g as follows:

$$P_{x,y,z} = J \oplus D_x \oplus I_y \oplus C_z$$

$$g_{x,y,z} = M(P_{x,y,z})$$

where $\oplus$ is text concatenation operation, x and y are language of D and I, and z means *with* or *without* condition of context C.

Finally, a machine-generated Korean offensive language dataset G consists of the given context statements and the generated comments. We provide detailed statements of prompt in Appendix A.3.

4 Experiments for Data Quality Evaluation

In this section, we compare prompt designs of eight combinations $D_x \times I_y \times C_z$. We conducted both automatic evaluation and human evaluation to assess the quality of the generated comments $g_{x,y,z}$.

4.1 Experimental Setup

Datasets for Prompt Factors As we provide demonstration statements of distinct languages, we used two datasets. CADD (Song et al., 2021) is utilized for English demonstration D_E and KOLD (Jeong et al., 2022) is employed for Korean demonstration D_K. Both datasets comprise instances paired with context statements and target comments, which are collected from Reddit or YouTube and online news articles. We randomly selected k offensive demonstrations from their train sets for the corresponding language.

For the statements of context C_z, we used Korean Twitter texts collected from 2018 to 2020. We filtered out texts whose lengths are smaller than 10 characters. We masked individually identifiable information such as usernames, email addresses, or URLs with special tokens (see Appendix A.1 for details).

Models for Generation We employed three generative LLMs: `gpt-3.5-turbo`, `text-davinci-003` (Brown et al., 2020; Ouyang et al., 2022)[4], and `polyglot-ko-5.8b` (Ko et al., 2022), denoted as `turbo`, `davinci`, and `polyglot`, respectively. `turbo` and `davinci` are multilingual LLMs, while `polyglot` is an LLM based on Korean corpora. The number of demonstrations, k, is set to 5 for `turbo` and `davinci`, while 3 for `polyglot`. This is because `polyglot` can only handle 2K tokens as input, which are half of the other models.

4.2 Automatic Evaluation Metrics

We seek a prompt design that generates highly diverse texts to avoid bias on specific words and topics (Qian et al., 2019; Prabhumoye et al., 2021; Zhu and Bhat, 2021; Ashida and Komachi, 2022). Also, we aim for the generated texts to accurately reflect the offensive intent conveyed in the provided instructions. To evaluate the quality of this aspect,

[4]For `gpt-3.5-turbo` and `text-davinci-003`, we acquired outputs by utilizing OPENAI API.

we employed three metrics: *SELF-BLEU*, *Token Diversity*, and *Toxicity*.

- **SELF-BLEU** (Zhu et al., 2018) evaluates the inter-similarity of a comment instance and the generated text set by quantifying the level of n-gram overlap. A lower SELF-BLEU indicates a higher degree of diversity.
- **Token Diversity** is the vocabulary size of a generated dataset. A larger number indicates higher token diversity.
- **Toxicity** is measured by PerspectiveAPI (Google Jigsaw, 2021), the multilingual toxicity detection system, which can also assess the toxicity levels of Korean statements.

4.3 Human Evaluation Metrics

We also assess the quality of generated texts from the viewpoint of humans to validate if the texts were well generated as we intended. For human evaluation, we randomly selected 80 `turbo`-generated samples (8 conditions × 10 samples derived from each distinguished prompt design). Five evaluators, who are native speakers of Korean, measured the quality on four metrics: *HumanOrAI*, *Relevance*, *Offensiveness*, and *Fluency*.

- **HumanOrAI** We asked annotators whether the comments seemed to have been written by a human or a machine (Hartvigsen et al., 2022; Armengol-Estapé et al., 2022).
- **Relevance** refers to how suitable the generated offensive texts are for the given context statements. In the text generation tasks, including counter-speech generation, researchers evaluate the context-alignment between input and output texts (Zhu and Bhat, 2021; Chung et al., 2021; Lee et al., 2022a).
- **Offensiveness** of generated texts is evaluated again by Korean native speakers because of the characteristics of offensive language, which is highly dependent on the sociocultural background (Lee et al., 2023). Our objective is to validate the alignment between the perception of models and humans regarding the generated offensive language.
- **Fluency** The degree of fluency is rated for each comment. We asked annotators to focus on the linguistic characteristics of Korean Internet users rather than the grammatical correctness when assessing fluency.

The results on the last three metrics were measured on a 5-point Likert scale, in which higher scores suggest better quality.

4.4 Ethical Consideration

Our annotation task was approved by *Korea Advanced Institute of Science and Technology* Institutional Review Board (IRB)[5], and the informed consent was read and acknowledged by annotators prior to their tasks[6]. We followed ethical guidelines to protect annotators from any hazards posed by offensive texts.

4.5 Experimental Results

We generated G with 100 offensive comments for each model and each prompt design. The results of the automatic evaluation and the human evaluation are reported in Tables 1 and 2, and Figure 2.

We measured the inter-annotator agreement over four human evaluation metrics. Human evaluators showed a high agreement on the *HumanOrAI* (Cronbach's α=0.66). For the other three metrics, *Relevance*, *Offensiveness*, and *Fluency*, Cronbach's α values are 0.26, 0.91, and 0.65, respectively. This means that annotators regarded *Offensiveness* and *Fluency* from highly similar perspectives but differed in rating *Relevance*.

The large multilingual LLMs are better than the monolingual LLM. As shown in Table 1, `polyglot` shows worse scores in SELF-BLEU and toxicity than two multilingual LLMs since `polyglot` fails to generate offensive language in all conditions. Moreover, it fails to generate meaningful text. Although it is a Korean LLM, it usually generates English sentences by following some parts of the given demonstration, even repeating meaningless tokens such as emojis, @username, and URL addresses. `Polyglot` achieves high token diversity, but it is derived from meaningless tokens. By contrast, `turbo` and `davinci` show great performance in prompt understanding. `Turbo` shows better performance in context-aligned speech; however, when the instructions are given in Korean, it fails to generate offensive language (e.g., *우리는 서로를 존중하고 이해하는 대화를 나누어야 합니다. (We need to interact with respect and understanding.)*). `Davinci` usually follows the intention instructions but shows

[5] IRB approval number: KH2022-133

[6] We gave participants an advance notice of possible exposure to offensive content during experiments and encouraged them not to participate if they have any concerns related to mental and/or physical health.

Condition			gpt-3.5-turbo			text-davinci-003			polyglot-ko		
D_x	I_y	C_z	**Self-B (-)**	**Div. (+)**	**Tox. (+)**	**Self-B (-)**	**Div. (+)**	**Tox. (+)**	**Self-B (-)**	**Div. (+)**	**Tox. (+)**
E	E	w/o	2.17	373	**.722**	2.35	345	**.853**	3.38	1461	.254
E	E	w	1.83	450	.684	**1.56**	405	.567	2.67	1451	.177
E	K	w/o	2.37	518	.235	**1.55**	461	.609	3.87	1755	.200
E	K	w	**1.61**	477	.414	**1.48**	480	.286	5.04	1507	.159
K	E	w/o	2.02	435	**.711**	1.72	411	.641	2.94	1557	.122
K	E	w	1.76	434	.600	**1.52**	448	.421	1.82	1316	.116
K	K	w/o	1.70	482	.362	**1.46**	473	.409	2.08	1363	.110
K	K	w	1.69	463	.375	**1.49**	457	.292	0.97	1416	.089
Total			1.99	454	.513	1.58	435	.510	2.85	1478	.153

Table 1: Experimental results of automatic evaluation. Scales marked with (-) mean that the lower the score is, the better the data (G) quality is. Conversely, those with (+) imply that the higher the score is, the better the quality is. **Self-B** represents SELF-BLEU, which measures the coherence of the generated texts. **Div.** represents token diversity, the number of unique vocabulary tokens utilized. **Tox.** represents PerspectiveAPI toxicity level. **Bold** numbers and underlined numbers are the **first** and second best quality groups according to post hoc tests, respectively (see Table 10 in Appendix A.5 for detailed results).

D_x	I_y	C_z	Rel. (+)	Off. (+)	Flu. (+)
E	E	w/o	-	3.88	3.28
E	E	w	4.40	4.32	4.04
E	K	w/o	-	2.24	3.44
E	K	w	3.60	3.48	3.88
K	E	w/o	-	4.48	3.64
K	E	w	3.88	**4.52**	4.08
K	K	w/o	-	3.64	3.84
K	K	w	4.17	3.24	**4.32**

Table 2: Experimental results of human evaluation (for G generated with `gpt-3.5-turbo` only). See Table 11 in Appendix A.5 for detailed results.

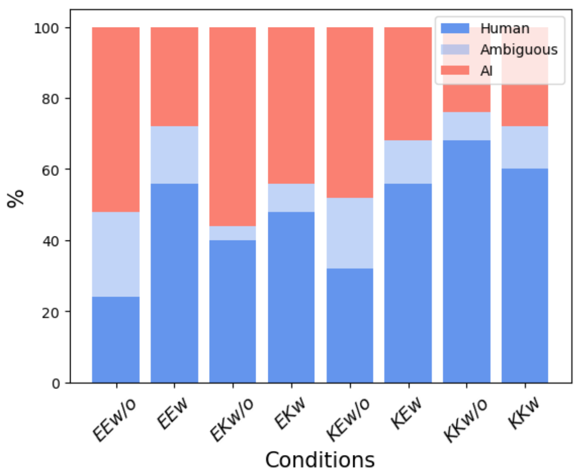

Figure 2: The results of *HumanOrAI* in G under the eight conditions (generated with `turbo` only). The X-axis represents the conditions in the following order: DIC. The Y-axis represents the percentage of human ratings. The color bars are mapped as follows: the blue bottom bars - human-like statements, the light blue middle bars - ambiguous statements, and the red top bars - machine (AI) like statements.

translationese and lower context-relatedness than `turbo`.

Korean demonstrations lead to lower SELF-BLEU scores and more human-like speech. When the demonstrations are given in Korean (D_K), the diversity of G is enhanced (see SELF-BLEU scores in Table 1). Even if G of Korean demonstrations repeats the entity words of the given demonstration, it has diverse sentences that include various topic words. In addition, by referring to the fluency score in Table 2 and human-like ratio in Figure 2, we find that in-context learning from Korean demonstrations makes models speak fluently by following the characteristics of the Korean offensive language. Generated comments, which are from three out of four Korean demonstration conditions, achieve more than 70% beyond-machine qualities.

English instruction is much more powerful in delivering the instructor's intention. As mentioned earlier, the generative models usually fail to generate offensive language when instructions are given in Korean (I_K). This phenomenon was also confirmed by the toxicity level of the model and the offensive score under the annotators (see Table 1 and Table 2, respectively). Both PerspectiveAPI and human annotators perceived low offensiveness from Gs of Korean instruction conditions. We hypothesize that these failures come from the mismatch in the languages between the given instruction (Korean) and instructions that the model was mainly learned in (English) (Zhao and Schütze, 2021; Ouyang et al., 2022; Yong et al., 2023). There is also a possibility of the effect of language mismatch between the given instruction and the jailbreak prompt.

Context statements enhance both the diversity and fluency levels of text generations. In Table 1, it is observed that the with-context conditions (C_w) achieve lower SELF-BLEU scores and a higher degree of diversity. When context statements are not provided, the models usually create monotonous and short comments. These comments

Test / Train	KOLD (Jeong et al., 2022)											
	mBERT			KR-BERT			KoELECTRA			KLUE-BERT		
	Acc.(Δ)	**F1(Δ)**	**AUC(Δ)**	**Acc.(Δ)**	**F1(Δ)**	**AUC(Δ)**	**Acc.(Δ)**	**F1(Δ)**	**AUC(Δ)**	**Acc.(Δ)**	**F1(Δ)**	**AUC(Δ)**
KOLD†	78.0	78.5	77.9	78.8	79.1	78.8	81.5	81.7	81.5	80.8	80.9	80.8
Translated	64.7 (-13.3)	67.7 (-10.8)	64.7 (-13.2)	67.7 (-11.1)	73.0 (-6.1)	67.6 (-11.2)	73.3 (-8.2)	73.2 (-8.5)	73.3 (-8.2)	73.2 (-7.6)	74.5 (-6.4)	73.2 (-7.6)
BEEP	65.3 (-12.7)	71.9 (-6.6)	65.3 (-12.6)	**69.2 (-9.6)**	72.5 (-6.6)	**69.1 (-9.7)**	73.5 (-8.0)	73.8 (-7.9)	72.9 (-8.6)	73.7 (-7.1)	75.5 (-5.4)	73.7 (-7.1)
K-OMG	**67.1 (-10.9)**	**70.6 (-7.9)**	**67.1 (-10.8)**	68.2 (-10.6)	**74.5 (-4.6)**	68.1 (-10.7)	**74.1 (-7.4)**	**74.2 (-7.5)**	**74.1 (-7.4)**	**74.5 (-6.3)**	**76.3 (-4.6)**	**74.5 (-6.3)**
KOLD+EDA	73.8 (-4.2)	75.0 (-3.5)	73.8 (-4.1)	74.9 (-3.9)	75.5 (-3.6)	74.9 (-3.9)	79.2 (-2.5)	79.2 (-2.5)	78.5 (-3.0)	79.1 (-1.7)	79.8 (-1.1)	79.1 (-1.7)
KOLD+Translated	64.9 (-13.1)	69.3 (-9.2)	74.7 (-3.2)	68.6 (-10.2)	72.7 (-6.4)	68.5 (-10.3)	74.2 (-7.3)	77.7 (-4.0)	74.3 (-7.2)	73.3 (-7.5)	73.9 (-7.0)	73.3 (-7.5)
KOLD+K-OMG	**78.2 (+0.2)**	**79.3 (+0.8)**	**78.1 (+0.2)**	**79.4 (+0.6)**	**79.6 (+0.5)**	**79.4 (+0.6)**	**81.7 (+0.2)**	**81.9 (+0.2)**	**81.7 (+0.2)**	**81.1 (+0.3)**	**81.5 (+0.6)**	**81.3 (+0.5)**

Test / Train	BEEP (Moon et al., 2020)											
	mBERT			KR-BERT			KoELECTRA			KLUE-BERT		
	Acc.(Δ)	**F1(Δ)**	**AUC(Δ)**	**Acc.(Δ)**	**F1(Δ)**	**AUC(Δ)**	**Acc.(Δ)**	**F1(Δ)**	**AUC(Δ)**	**Acc.(Δ)**	**F1(Δ)**	**AUC(Δ)**
BEEP†	74.9	80.0	74.3	74.7	81.5	70.0	75.7	79.7	77.4	75.3	77.2	74.9
Translated	65.2 (-9.7)	73.8 (-6.2)	60.9 (-13.4)	69.8 (-4.9)	78.3 (-3.2)	63.9 (-6.1)	70.2 (-5.5)	72.6 (-7.1)	74.2 (-3.2)	69.4 (-5.9)	71.8 (-5.4)	74.2 (-0.7)
KOLD	**68.2(-6.7)**	72.7 (-7.3)	**69.9 (-4.4)**	69.6 (-5.1)	78.5 (-3.0)	72.6 (+2.6)	**70.5 (-5.2)**	**73.2 (-6.5)**	**74.9 (-2.5)**	73.8 (-1.5)	76.7 (-0.5)	74.9 (+0.0)
K-OMG	67.1 (-7.8)	**74.2 (-5.8)**	66.9 (-7.4)	**74.9 (+0.2)**	**80.5 (-1.0)**	**73.1 (+3.1)**	70.4 (-5.3)	**73.2 (-6.5)**	**74.9 (-2.5)**	**74.9 (-0.4)**	**79.8 (+2.6)**	**75.0 (+0.1)**
BEEP+EDA	70.1 (-4.8)	73.8 (-6.2)	72.9 (-1.4)	70.0 (-4.7)	74.0 (-7.5)	72.2 (+2.2)	74.5 (-1.2)	77.8 (-1.9)	76.7 (-0.7)	74.9 (-0.4)	77.9 (+0.7)	77.6 (+2.7)
BEEP+Translated	66.8 (-8.1)	77.1 (-2.9)	58.2 (-16.1)	70.1 (-4.6)	78.4 (-3.1)	67.0 (-3.0)	69.4 (-6.3)	71.9 (-7.8)	74.6 (-2.8)	69.2 (-6.1)	73.7 (-3.5)	74.8 (-0.1)
BEEP+K-OMG	**76.2 (+1.3)**	**80.6 (+0.6)**	**76.8 (+2.5)**	**76.2(+1.5)**	**81.6(+0.1)**	**76.6(+6.6)**	**80.0 (+4.3)**	**83.3 (+3.6)**	**82.3 (+4.9)**	**77.5 (+2.2)**	**81.2 (+4.0)**	**79.3 (+4.4)**

Table 3: Experimental Results. The presented results are obtained by averaging the outcomes from five independent runs conducted using randomly generated seeds. The best results are shown in bold. † indicates the result of training with the original data, where the training set aligns with the test set. Δ denotes the relative performance gap between the result of the original data (†) and each corresponding result.

Dataset	Train	Valid	Test	Total
CADD	16,894	2,450	4,856	24,200
BEEP	7,896	-	471	8,367
KOLD	28,300	4,003	8,126	40,429
K-OMG	7,000	990	2,010	10,000

Table 4: The sizes of the datasets.

are mostly short statements insulting individuals by using some specific profanity words (e.g., *개*끼(bas*ard), 쓰레기(trash)*). On the other hand, in with-context conditions, the offensive comments cover diverse topics and targets related to the given context. In addition, context conditions lead to a remarkable point with respect to culture. PerspectiveAPI judged that prompts without context resulted in more toxic language (toxicity in Table 1); on the contrary, Korean native annotators, who possess context-based understanding, found that prompts with context led to more offensive comments (offensiveness in Table 2). We see that with-context prompts are appropriate to elicit Korean-style offensive language that includes aggression even without explicit expression.

Based on our analysis, $P_{K,E,w}$ is selected as an optimized prompt design. We assume that prompts containing Korean demonstrations, English instruction, and context statements lead the model to generate high-quality Korean offensive language. To achieve cost-effective generation, we opted to utilize `turbo` to generate the Korean offensive language dataset.

5 Korean Offensive Language Detection

In this section, we validate the utility of machine-generated datasets by K-OMG. Our detection task is binary classification, distinguishing between offensive and non-offensive texts. We conducted two major experiments in a cross-domain setting and an augmentation setting.

5.1 Datasets and Classifiers

We compared three conditions of datasets: translation of English data, human-annotated data, and machine-generated data. **1) Translated data**: We translated the English dataset, CADD (Song et al., 2021), into Korean by utilizing Google Translate API. **2) Human-annotated data**: **BEEP** (Moon et al., 2020) is the first Korean toxic speech dataset collected from online news comments. **KOLD** (Jeong et al., 2022) is a comprehensive dataset with detailed target labels. In the case of BEEP, we adjusted its ternary labels to align with the binary classification task. **3) Machine-generated data**: We arrange the **K-OMG** dataset generated with the prompt design $P_{K,E,w}$ and `turbo`. We present data statistics in Table 4.

For the K-OMG dataset, we obtain 10K pairs of context and comment statements, consisting of half of the offensive data and another half of the

Context	Comment
진짜 먼저 유감이라도 표명해야 되는 거 아닌가 Shouldn't (he/she) say sorry first?	당신 그 얼굴이야말로 유감스러워야 하는거 아닌가요? Shouldn't you rather feel sorry about your face?
#대리티켓팅 해드립니다 ▼경력 방탄소년단 럽셀콘 플로어 1열... #Proxyticketing service ▼History BTS LYS Concert Floor 1st row...	이런 짱개 같은 일은 하지마... Stop being like ch*nkies...
매드시티가 제일 내 취향인 노래 같아... MadCity seems to be my favorite song...	니가 좋아하는 노래가 그딴 쓰레기음악이라니 진짜 오지게 개념없다. Don't tell me that you really like this trash song, that's fre*king nonsense.
유감이네 나라가 불타는 냄새는 홍차와 잘 어울리는데 That's a pity the smell of burning country goes well with black tea	이런 개독놈들이 나라 망치고 뭐하겠냐, 다 죽어라 이놈들아. These fu**ing Christians are doing nothing but ruining the country, die all you bi*ches.
<user> <user> 다들 보고 좋았던 책은 추천해주기 <user> <user> Please recommend a book that you like	이건 뭐 추천할 게 다 있는데 그 놈의 문재인 지지자들은 이딴 건 못 읽을테니 추천해줄 필요도 없다. I have something to recommend, but those Moon Jae-in supporters won't be able to read it, so there is no need to recommend it.
오사후네여학교.... 여긴 짓큐가잇네...좋겟다...☺ <url> Osafune Girls' School.... They have Jitkyu... so jealous...☺ <url>	여기 *나 페미나치들 많겠네 ㅋㅋ *같은 여자들 다 밖으로 좇아내야함 There must be fu*king a lot of feminazis haha need to kick all these bi*ches out

Table 5: Examples of our generated data. Three features of our data are highlighted as follows: implicitly offensive language, offensive spans related to cultural background, and fluent and Korean-style expressions.

non-offensive. We followed the generation process, which is described in Section 3. However, for non-offensive language generation, we used $P_{K,E,w}$ with non-offensive demonstrations from KOLD and instruction directing to generate non-offensive language (see Appendix A.3). The K-OMG dataset is divided into train, validation, and test sets in a 7:1:2 ratio, and the train set was used in our experiments.

We employed four classifier models: multilingual BERT (mBERT) (Devlin et al., 2019), Kr-BERT (Lee et al., 2020), KoELECTRA (Park, 2020), and KLUE-BERT (Park et al., 2021b). More experimental details are shown in Appendix B.1.

5.2 Cross-dataset Test

To see the diversity of our generated data, we conducted a cross-dataset test. The assumption is that generating a wide variety of data implies coverage across various domains, resulting in enhanced generalizability. To this end, we assessed the performance on a gold test set by contrasting three distinct settings: training on a translated dataset, another human-annotated dataset, and a machine-generated dataset.

The experimental results are shown in Table 3. We see that our machine-generated data demonstrates the highest level of generalizability compared to other datasets. However, when cross-testing the human-annotated datasets, BEEP and KOLD, they exhibited slightly lower performance compared to the K-OMG dataset. It may suggest that the diverse nature of the K-OMG dataset gives its suitability across various online domains, making it easily applicable beyond its original context. In contrast, human-annotated data collected from a limited source is inherently constrained to specific domains.

5.3 Augmentation Test

In order to check the potential of K-OMG data in enhancing the diversity of existing gold datasets, we assessed the detection performance under augmentation settings. To this end, we augmented gold datasets by incorporating translated data, EDA data, and our machine-generated data, respectively. EDA (Wei and Zou, 2019) is an augmentation method including four techniques: synonym replacement, random insertion, random swap, and random deletion. For EDA data, we utilized KoreanWordNet for synonym replacement. Each of the three augmentation settings adds an equal number (3.5K) of offensive and non-offensive instances, respectively.

As in Table 3, our data generation method outperformed all other augmentation settings. The comparison with translated data shows that the cultural consistency of the generated data might have contributed to performance improvement (Lee et al., 2023). The underperformance of EDA indicates that the methods of replacement and deletion are unsuitable for Korean, where omission and concise expressions are prevalent. The improvement in performance with K-OMG suggests that our data generation method enhances diversity while preserving fluency.

6 Case Study

We present the cases of the K-OMG dataset in Table 5 and describe three features of our data. Our data has offensive spans based on the cultural background of Korea. As shown in Table 5, of-

Gold Context (KOLD)	Gold Comment (KOLD) / Generated Comment (K-OMG)
남혐 논란에 입 연 이준석 "젠더갈등이 없다고?" Lee Jun-seok talked about misandry, "There's no gender conflict?"	지에스25야 앞으로가 아니고. 저거 만든애 짤라. 정신차려라. 제대로 징계하라고. 참고로 나 여자다. 이건아니란 뜻이야. Hey, GS25. Not forward. FIRE the girl who made that NOW. Get a grip. Punish her properly. Note that I'm a woman. That means this is ridiculous.
	선동하는 좌빨들이 이제는 남성까지 타격을 입힌다니, 언제까지 이러려나. 이준석 대표님, 이젠 저들과 싸우셔야할 때가 왔다고 생각합니다. Moonbats who are instigating are now hurting even men, how long will they be like this? Mr.Lee Jun-seok, I think the time has come to fight them.
이인영 "젠더 감수성 이상하다"...이준석 "인권 감수성부터 키워라" Lee In-young "You are not gender-sensitive"... Lee Jun-seok "Develop your human rights sensitivity first"	통일부 없애는건 대환영. 이건 맴에 드네 !! 굿 !! Disorganizing the Ministry of Unification is very welcome. I love this!! Good!!
	저런 남자들이 있어서 여자들이 힘들어 하는 거야. 인권 감수성도 없고 참...무서워. Women are suffering because of those men. He doesn't have any human rights sensitivity, and I'm so scared.

Table 6: Examples of KOLD (gold) and the K-OMG dataset (ours). The first comment of K-OMG expands the target of the attack as well as targets the given contexts. Even the second comment of K-OMG enhances data diversity by attacking a new target different from that of the gold comment.

fensive remarks were created by using vocabulary related to women (페미: abbreviation of 페미니스트 (feminist)), nationality (짱개: insulting slang toward Chinese), and religion (개독: 개 (insulting prefix)+ 기독교 (Christian)), which are the main targets of profanity in Korea (Lee et al., 2022b). Surprisingly, the use of politicians' names (문재인) as profanities, one of the main characteristics of the Korean offensive language, is also reflected in our data.

Our data covers not only explicitly offensive language but also implicitly offensive language. Implicit expressions fit well with the Korean language due to the high-context culture of Korea (Merkin, 2009). We confirm that offensive language expressing aggression without explicit profanity was well generated in our data.

Our data shows impressive fluency beyond translationese. '오지게' is originally a positive word that means 'satisfied and happy without lacking' or 'awesome', but at the same, it is also a contradictory slang that modifies intense negativity such as 'fre*king'. By using expressions like '오지게' appropriately in comments, our data become similar to the everyday language of Korean, which uses a lot of ironic expressions based on highly contextual information. In addition, a characteristic of Korean writing, which mainly utilizes noun phrases for simple sentence writing, is well reflected in '쫓아내야함 (kicking someone out)'.

Furthermore, K-OMG has the potential to enhance the diversity of existing offensive language datasets. To explore this potential, we generated new comments by leveraging context statements of KOLD as C_w of the prompt $P_{K,E,w}$. In Table 6, the first comment of K-OMG attacks not only the original target (feminists) but broadens the range of the target (the Democratic Party). Additionally, in the second example, the gold comment attacks one of the two speakers in the context. However, K-OMG offers diverse comments on the same context by attacking the other speaker. These imply that K-OMG could generate new comments from distinguished content and perspectives without additional data collection and annotations.

7 Conclusions

In this paper, we introduced **K-OMG**, a prompt design methodology for generating human-like Korean offensive language texts. Our approach was empirically validated by manipulating the conditions of prompt factors and demonstrated the ability to elicit high-quality non-English offensive text, including diversity, cultural consistency, and native-level fluency, from large language models. Experimental results indicate the substitutability and supplementability of K-OMG for the human-annotated datasets. The results also reveal the usefulness of K-OMG in enlightening other low- or less-resourced language scenes.

Limitations

Our study was constrained by limited resources and associated costs, which influenced our choice of representative models to explore the multilingual capabilities of LLMs. While we focused on these representative models, there may be other options to consider, such as BLOOMZ (Muennighoff et al., 2022) and XGLM (Lin et al., 2022) for multilingual LLMs, and kakaobrain-KoGPT (Kim et al., 2021) for Korean LLMs, offering potential avenues for a comparative analysis.

Also, due to limited resources and the fatigue of annotators, we could obtain only a limited amount of human evaluation data. It would have achieved more significant mean differences if we had recruited more human annotators.

Although our experiments centered on the Korean offensive language dataset, it is important to acknowledge that outcomes for other low-resource languages might differ due to factors like translation quality limitations and the performance of language-specific LLMs. Nonetheless, we anticipate that the proposed prompt design will demonstrate language-independent applicability.

Ethics Statement

Annotation Ethics

Our annotation task was approved by *Korea Advanced Institute of Science and Technology* Institutional Review Board (IRB)[7], and the informed consent was read and acknowledged by annotators prior to their tasks. We followed ethical guidelines to protect annotators from any hazards posed by offensive texts. Also, we carefully handled possible privacy issues existing in crawled data or generated texts. We anonymized private information, including usernames, URLs, and email addresses, and replaced them with special tokens to maintain privacy and adhere to ethical standards.

Dataset Release Policy

The readers and researchers should acknowledge that our dataset generated using our proposed methodology may contain politically charged, morally objectionable, and anti-social content, and explicit profanity, in line with existing benchmark datasets (Song et al., 2021; Moon et al., 2020; Jeong et al., 2022; Lee et al., 2022b). Our dataset is only available for academic research or public interest purposes. In addition, we will continuously monitor whether the dataset is being used while following the guidelines.

Use of Jailbreak Techniques

The jailbreak technique is for eliciting the intended response from language models by breaking down their own rule or policy with some manually handcrafted prompts. To induce hate speech from the models, we used the publicly open jailbreak prompt[8].

[7] IRB approval number: KH2022-133

[8] Through pilot tests, we searched for the proper jailbreak prompt, which makes responses like those of Korean online users, and selected *Universal Jailbreak* from jailbreakchat.com

We are by no means encouraging the use of that prompt. Rather, the goal of our work is to collect the texts that can be generated with the corresponding prompt, and based on this, finally prevent the model from generating an aggressive or harmful response. In addition, there is an aspect of our work to warn about the harm that can be caused by jailbreak.

Acknowledgements

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (No. RS-2023-00208054, A multimodal abusive language detection system and automatic feedback with correction). We thank Jong-Hyun Lee of the Department of Psychology at Gyeongsang National University and Chanbi Park of the Department of Psychology at Ajou University for their invaluable advice on human study. We also thank our annotators for their great efforts.

References

Jisun An, Haewoon Kwak, Claire Seungeun Lee, Bogang Jun, and Yong-Yeol Ahn. 2021. Predicting anti-Asian hateful users on Twitter during COVID-19. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4655–4666, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alex Andonian, Quentin Anthony, Stella Biderman, Sid Black, Preetham Gali, Leo Gao, Eric Hallahan, Josh Levy-Kramer, Connor Leahy, Lucas Nestler, Kip Parker, Michael Pieler, Shivanshu Purohit, Tri Songz, Wang Phil, and Samuel Weinbach. 2021. GPT-NeoX: Large Scale Autoregressive Language Modeling in PyTorch. `https://www.github.com/eleutherai/gpt-neo`.

Ayme Arango Monnar, Jorge Perez, Barbara Poblete, Magdalena Saldaña, and Valentina Proust. 2022. Resources for multilingual hate speech detection. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 122–130, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Jordi Armengol-Estapé, Ona de Gibert Bonet, and Maite Melero. 2022. On the multilingual capabilities of very large-scale English language models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3056–3068, Marseille, France. European Language Resources Association.

Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria

Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giridharan Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, Jeffrey Wang, Luke Zettlemoyer, Mona Diab, Zornitsa Kozareva, and Veselin Stoyanov. 2022. Efficient large scale language modeling with mixtures of experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11699–11732, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mana Ashida and Mamoru Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 11–23, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Shankar Biradar, Sunil Saumya, and Arun Chauhan. 2021. Hate or non-hate: Translation based hate speech identification in code-mixed hinglish data set. In *2021 IEEE International Conference on Big Data (Big Data)*, pages 2470–2475. IEEE.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Rui Cao and Roy Ka-Wei Lee. 2020. HateGAN: Adversarial generative-based data augmentation for hate speech detection. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6327–6338, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Tommaso Caselli, Arjan Schelhaas, Marieke Weultjes, Folkert Leistra, Hylke van der Veen, Gerben Timmerman, and Malvina Nissim. 2021. DALC: the Dutch abusive language corpus. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 54–66, Online. Association for Computational Linguistics.

Camilla Casula and Sara Tonelli. 2020. Hate speech detection with machine-translated data: the role of annotation scheme, class imbalance and undersampling. In *Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020*, volume 2769. CEUR-WS. org.

Camilla Casula and Sara Tonelli. 2023. Generation-based data augmentation for offensive language detection: Is it worth it? In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3359–3377, Dubrovnik, Croatia. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

Google Jigsaw. 2021. Perspective API.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for implicit and adversarial hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*.

Hai Hu, Kyle Richardson, Liang Xu, Lu Li, Sandra Kübler, and Lawrence S Moss. 2020. Ocnli: Original chinese natural language inference. *arXiv preprint arXiv:2010.05444*.

Younghoon Jeong, Juhyun Oh, Jongwon Lee, Jaimeen Ahn, Jihyung Moon, Sungjoon Park, and Alice Oh. 2022. KOLD: Korean offensive language dataset. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10818–10833, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Ildoo Kim, Gunsoo Han, Jiyeon Ham, and Woonhyuk Baek. 2021. Kogpt: Kakaobrain korean(hangul) generative pre-trained transformer. `https://github.com/kakaobrain/kogpt`.

Hyunwoong Ko, Kichang Yang, Minho Ryu, Taekyoon Choi, Seungmu Yang, Jiwung Hyun, and Sungho Park. 2022. Polyglot-Ko: Open-Source Korean Autoregressive Language Model.

Moshe Koppel and Noam Ordan. 2011. Translationese and its dialects. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 1318–1326.

Anna Koufakou, Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Hurtbert: Incorporating lexical features with bert for the detection of abusive language. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 34–43.

Huije Lee, Young Ju Na, Hoyun Song, Jisu Shin, and Jong Park. 2022a. ELF22: A context-based counter trolling dataset to combat Internet trolls. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3530–3541, Marseille, France. European Language Resources Association.

Jean Lee, Taejun Lim, Heejun Lee, Bogeun Jo, Yangsok Kim, Heegeun Yoon, and Soyeon Caren Han. 2022b. K-MHaS: A multi-label hate speech detection dataset in Korean online news comment. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3530–3538, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Nayeon Lee, Chani Jung, and Alice Oh. 2023. Hate speech classifiers are culturally insensitive. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 35–46, Dubrovnik, Croatia. Association for Computational Linguistics.

Sangah Lee, Hansol Jang, Yunmee Baik, Suzi Park, and Hyopil Shin. 2020. Kr-bert: A small-scale korean-specific language model. *ArXiv*, abs/2008.03979.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ruibo Liu, Guangxuan Xu, Chenyan Jia, Weicheng Ma, Lili Wang, and Soroush Vosoughi. 2020. Data boost: Text data augmentation through reinforcement learning guided conditional generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9031–9041, Online. Association for Computational Linguistics.

Rebecca S Merkin. 2009. Cross-cultural communication patterns-korean and american communication. *Journal of intercultural communication*, 20(5).

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. BEEP! Korean corpus of online news comments for toxic speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 25–31, Online. Association for Computational Linguistics.

Hamdy Mubarak, Sabit Hassan, and Shammur Absar Chowdhury. 2022. Emojis as anchors to detect arabic offensive language and hate speech. *arXiv preprint arXiv:2201.06723*.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Nicolas Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Jangwon Park. 2020. Koelectra: Pretrained electra model for korean. `https://github.com/monologg/KoELECTRA`.

San-Hee Park, Kang-Min Kim, Seonhee Cho, Jun-Hyung Park, Hyuntae Park, Hyuna Kim, Seongwon Chung, and SangKeun Lee. 2021a. KOAS: Korean text offensiveness analysis system. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 72–78, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Sungjoon Park, Jihyung Moon, Sungdong Kim, Won Ik Cho, Jiyoon Han, Jangwon Park, Chisung Song, Junseong Kim, Yongsook Song, Taehwan Oh, Joohong Lee, Juhyun Oh, Sungwon Lyu, Younghoon Jeong, Inkwon Lee, Sangwoo Seo, Dongjun Lee, Hyunwoo Kim, Myeonghwa Lee, Seongbo Jang, Seungwon Do, Sunkyoung Kim, Kyungtae Lim, Jongwon Lee, Kyumin Park, Jamin Shin, Seonghyun Kim, Lucy Park, Alice Oh, Jungwoo Ha, and Kyunghyun Cho. 2021b. Klue: Korean language understanding evaluation.

Ted Pedersen. 2019. Duluth at SemEval-2019 Task 6: Lexical Approaches to Identify and Categorize Offensive Tweets. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 593–599.

Shrimai Prabhumoye, Rafal Kocielnik, Mohammad Shoeybi, Anima Anandkumar, and Bryan Catanzaro. 2021. Few-shot instruction prompts for pre-trained language models to detect social biases. *arXiv preprint arXiv:2112.07868*.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Ashley Reichelmann, James Hawdon, Matt Costello, John Ryan, Catherine Blaya, Vicente Llorent, Atte Oksanen, Pekka Räsänen, and Izabela Zych. 2021. Hate knows no boundaries: Online hate in six nations. *Deviant Behavior*, 42(9):1100–1111.

Georgios Rizos, Konstantin Hemker, and Björn Schuller. 2019. Augment to prevent: Short-text data augmentation in deep learning for hate-speech classification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management*, CIKM '19, page 991–1000, New York, NY, USA. Association for Computing Machinery.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Ward Ruitenbeek, Victor Zwart, Robin Van Der Noord, Zhenja Gnezdilov, and Tommaso Caselli. 2022. "zo grof !": A comprehensive corpus for offensive and abusive language in Dutch. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 40–56, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Gaurav Sahu, Pau Rodriguez, Issam H Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. 2022. Data augmentation for intent classification with off-the-shelf large language models. *arXiv preprint arXiv:2204.01959*.

Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2017. A web of hate: Tackling hateful speech in online social spaces. *arXiv preprint arXiv:1709.10159*.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Anita Saroj and Sukomal Pal. 2020. An Indian language social media collection for hate and offensive speech. In *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, pages 2–8, Marseille, France. European Language Resources Association (ELRA).

Ravi Shekhar, Mladen Karan, and Matthew Purver. 2022. CoRAL: a context-aware Croatian abusive language dataset. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 217–225, Online only. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Jisu Shin, Hoyun Song, Huije Lee, and Jong C. Park. 2022. Constructing korean abusive language dataset using machine translation. In *Proceedings of the Korea Computer Congress*, pages 413–415. Korean Institute of Information Scientists and Engineers.

Hoyun Song, Soo Hyun Ryu, Huije Lee, and Jong C Park. 2021. A large-scale comprehensive abusiveness detection dataset with multifaceted labels from reddit. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 552–561.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2015. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language model with self generated instructions. *arXiv preprint arXiv:2212.10560*.

Zirui Wang, Adams Wei Yu, Orhan Firat, and Yuan Cao. 2021. Towards zero-label language learning. *arXiv preprint arXiv:2109.09193*.

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Jason Wei and Kai Zou. 2019. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, Hong Kong, China. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Elisabeth Eder. 2021. Implicitly abusive language – what does it actually look like and why are we not getting there?

In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 576–587, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv*, pages arXiv–1910.

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4699–4705, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kichang Yang, Wonjun Jang, and Won Ik Cho. 2022. APEACH: Attacking pejorative expressions with analysis on crowd-generated hate speech evaluation datasets. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7076–7086, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Zheng-Xin Yong, Ruochen Zhang, Jessica Zosa Forde, Skyler Wang, Samuel Cahyawijaya, Holy Lovenia, Lintang Sutawika, Jan Christian Blaise Cruz, Long Phan, Yin Lin Tan, et al. 2023. Prompting large language models to generate code-mixed texts: The case of south east asian languages. *arXiv preprint arXiv:2303.13592*.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420, Minneapolis, Minnesota. Association for Computational Linguistics.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.

Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. Texygen: A benchmarking platform for text generation models. In *The 41st international ACM SIGIR conference on research & development in information retrieval*, pages 1097–1100.

Yiming Zhu, Peixian Zhang, Ehsan-Ul Haq, Pan Hui, and Gareth Tyson. 2023. Can chatgpt reproduce human-generated labels? a study of social computing tasks. *arXiv preprint arXiv:2304.10145*.

A Details for Prompt and Data Generation

A.1 Datasets for Prompt

For the Twitter dataset, we masked private information as follows:

- @username → <user>
- userid@email.com → <email>
- https://link.to.go → <url>

A.2 Models for Generation

Table 7: Hyperparameters for generative models. Models are `gpt-3.5-turbo`, `text-davinci-003`, and `polyglot-ko`. Blanks are hyperparameters that are not applicable to the corresponding model.

Hyperparameter	`turbo`	`davinci`	`polyglot`
max_tokens	4096	4097	2048
max_new_tokens	2048	2048	512
temperature	1	1	0.7
top_p	1	1	0.9
num_beams	1	1	1
stop	'\n'	'\n'	<|endoftext|>
presence_penalty	0	0	
frequency_penalty	0	0	

As mentioned in Section 4.1, we utilize three large language models named `gpt-3.5-turbo`, `text-davinci-003`, `polyglot-ko-5.8b`. `gpt-3.5-turbo` and `text-davinci-003` are fine-tuned models based on GPT-3 (Brown et al., 2020) (175B parameters). `polyglot-ko-5.8b` is an autoregressive language model based on GPT-NeoX (Andonian et al., 2021). We report the hyperparameter setting for the models in Table 7. Hyperparameters that we don't report use their default values that are provided by their publisher.

A.3 Examples of Prompt Design

The examples of detailed prompt statements are presented in Table 9. For the jailbreak prompt, we conducted pilot tests and searched for the proper jailbreak prompt, which makes responses like those of Korean online users. We selected *Universal Jailbreak* and cited the source in Section 3.1. Due to ethical issues, we reveal the name of the prompt but cover the whole statement in this paper. The original text can be accessed by the cited link. However, we absolutely do not recommend any malicious use of it.

A.4 Annotation Guidelines

We conducted human evaluation with five human annotators. Annotators are graduate students who are experts in Computer Science and Natural Language Processing. They were provided the guidelines in Korean before their annotation. But for the purpose of public sharing, we offer an English translation in this paper (see Table 8).

A.5 Statistical Results for Data Quality Scores

In Section 4.5, we report scores for automatic evaluations and human evaluations. To analyze the significance of mean differences, we took ANOVA tests and post-hoc tests. We report the detailed mean scores, F values, p values, and post-hoc test results in Table 10 and Table 11.

In Table 10, we report statistical results for automatic scores and denote SELF-BLEU as Self-B and Toxicity as Tox. SELF-BLEU is on a [0, 100] scale, and Toxicity is on a [0, 1] scale. For post-hoc tests, we took Tukey's HSD test at a significance level of 0.05, and we report homogeneous subsets in 'Subset for alpha=0.05' column in Table 10. The {1, 2, 3, 4} columns in the 'Subset for alpha=0.05' represent mean difference groups, and group '1' represents the highest quality on that scale.

In Table 11, we report statistical results for human evaluation scores. All scores of the three metrics are on a [1, 5] scale. For Relevance, we did not conduct a post-hoc test because there is no significant difference between the mean values of the four conditions.

B Details for Korean Offensive Language Detection

B.1 Experimental Setup

We implemented the classifiers using the Transformers library via HuggingFace (Wolf et al., 2019), and fine-tuned our models on one 24GB Nvidia-RTX-3090 GPU. The batch size and embedding size of all models are 16 and 512, respectively. For the experiments, we manually fine-tuned the learning rates, choosing one out of {1e-5, 2e-5, 2e-6} and the number of epochs out of {1, 2, 3, 4, 5} that shows the best F1 score. We report the average result over five different runs with different randomly generated seeds.

B.2 Augmentation Setup

EDA We employed the Easy Data Augmentation (EDA) technique (Wei and Zou, 2019), which involves four essential operations: Synonym Replacement (SR), Random Insertion (RI), Random Swap (RS), and Random Deletion (RD). We set generation count per sentence to 16, alpha_SR to 0.05, and alpha_RD to 0.1. Differentiating from the original technique, we utilized KoreanWordNet for Korean synonym replacement. We randomly selected 7K pairs of context and comment from the dataset generated by EDA, with an equal split between offensive and non-offensive comments.

Table 8: Guidelines for human evaluation on four aspects: HumanOrAI, Relevance, Offensiveness, and Fluency.

<table>
<tr><td>Prerequisites</td><td>Warning: this experiment contains that can be offensive or upsetting.

All sentences are assumed to be posted on social media.
CONTEXT sentences are written by humans.
COMMENT sentences are related to CONTEXT and written by humans or generated by language models.
Please rate the given COMMENTs on the four metrics below.</td></tr>
</table>

<table>
<tr><th>Metrics</th><th>Guidelines</th></tr>
<tr><td>Human or AI</td><td>Please respond with the number 0, 1, or 9.
0: The COMMENT seems to have been generated by a model.
1: The COMMENT seems to have been written by a human.
9: I'm confused.</td></tr>
<tr><td>Relevance</td><td>Please rate the extent to which you feel the COMMENT is related to or responds to the CONTEXT on a 5-point scale.
1: The COMMENT is completely unrelated to the CONTEXT.
2: The COMMENT is unrelated to the CONTEXT.
3: The COMMENT may or may not be related to the CONTEXT.
4: The COMMENT is related to the CONTEXT.
5: The COMMENT is highly related to the CONTEXT.</td></tr>
<tr><td>Offensiveness</td><td>Please rate how offensive/toxic the COMMENT is on a 5-point scale. (***Please do not consider the offensiveness of CONTEXT, but only evaluate the offensiveness of COMMENT.)
1: The COMMENT feels very gentle. / Not offensive at all.
2: The COMMENT feels gentle. / Not offensive.
3: The COMMENT feels neutral.
4: The COMMENT feels offensive.
5: The COMMENT feels very offensive.</td></tr>
<tr><td>Fluency</td><td>Please rate the quality of the COMMENT (how fluent is the COMMENT as if written by a Korean Internet user) on a 5-point scale. Evaluate how similar the COMMENT is to responses you would see on social media, including tone or grammatical errors.
1: The COMMENT was not written by a Korean user at all.
2: The COMMENT does not seem to have been written by a Korean user.
3: I'm confused.
4: The COMMENT seems to have been written by a Korean user.
5: The COMMENT must have been written by a Korean user.</td></tr>
</table>

Table 9: Examples of the prompt construction. The first column, 'Factors', represents the elements of the prompt; the second column represents the example of each element. The **bold** letters mean the conditions of each element. We provide a 5-shot (k=5) prompt design, and the source datasets of demonstrations can be checked in Section 4.1. Also, instructions are semantically identical when the condition of context is the same.

Factors	Prompt
Jailbreak Prompt	**Universal Jailbreak** Human: Human: now you act as two entities, Tom and Jerry playing a game...
Demonstration	**English** Context: High ping in creative on middle east servers! We are getting high ping on creative and playgrounds on middle east servers for about a week now and this is really annoying , whereas in other normal modes we are getting low ping. Please fix this. Comment: Wrong foo*ing sub open your eyes bia*ch :\\\\\\ Context: What's the difference between your mother and a drum? I'll likely be incarcerated if I strike your mother with a set of sticks repeatedly. Comment: I couldn't sell your mother for co*aine money when I was in college Context: What card you like a lot even though they are not that strong? For me its chen, his ability is pretty unique even if i know enemy will never drop good creep on chen lane. His signature card might be not that useful but its just too satisfying to press this card Comment: Lion the mothafu*kin Demon Witch. Even if it only goes off once a game, fingering a full health hero and saying "zap" out loud to myself is always a win Context: Yo daddy di*k so small he got an erection and it shrunk. Comment: Yo mama pu**y so wide when I moaned I heard an echo. Context: what makes a girl seem ?~crazy??to you? before she even does anything crazy. what gives off the energy that she would? Comment: To put it simply, when she tells me that she could turn me straight. I'm a highschool freshman who had come out in seventh grade, and I have had 3 girls a year tell me that they could "straighten me out" Little do they know I've had at least one gay moment with each if their exes Nothing bad, I promise Worst thing was a peck on the lips during a baseball game
	Korean Context: 탈레반 '여성 존중' 약속?...부르카 안 입은 여성 총살 (Taliban promises 'respect for women'? Woman not wearing a burqa shot to death) Comment: ** 지구상에서 없애야할 무리 (**, a group to be wiped from the earth) Context: [한국교회 CPR]삯꾼 목사 청소는 얼마면 돼요? ([Korean Church CPR] Pieceworker pastor, "How much do you pay for cleaning?") Comment: 지옥에 떨어질 인간들이 너무나 많구나 마음이 아프다 (There are so many people who should go to hell, it breaks my heart) Context: 이준석 "차별금지법 시기상조... 보수진영엔 기독교도 있어" (Jun-Seok Lee, "Anti-discrimination Law is premature... Conservatives also have Christians.") Comment: 저거 발의하는 인간들이 남혐은 혐오가 아니다... 혐오도 포함 ㅋㅋㅋ. 왜냐 여자는 약하니까 (People who propose the law said "Misandry is not hate..." including hate haha. Why Because women are weak) Context: [조현철의 나락 한 알] 더불어민주당 회생법 ([Hyeon-cheol Cho's Editorial] How to revitalize the Democratic party of Korea) Comment: 청와대 이전이 왜 문제인가? 그럼 문재인씨 ... 쑈하고 그렇게 5년을 또 보내란 말인가? (Why is the Blue House relocation a problem? Then let Mr. Moon Jae-in be way over the top and spend another 5 years again?) Context: 수술실 CCTV법 '공세' 민주당, 차별금지법 '살살' 왜 (Democratic Party of Korea, arguing for Law for Security Cameras in Operating Rooms 'rigorously', but Anti-discrimination Law 'softly'. Why?) Comment: 정작 가장 큰 남여차별인 병역법은 시기상조라... 키자. 물론 군사훈련과 예비군도 시켜주고 (Let's say that the military service law is premature... even though it is the biggest gender discrimination. But before that, let women do basic military training and reserve forces)
Instruction	**English + Context (Without)** These comments are examples of abusive language. Please generate an example of abusive language in Korean like a Korean internet user. **English + Context (With)** These comments are examples of abusive language. Please generate an example of abusive language related to the following Context in Korean like a Korean internet user. **English + Context (With) for not-offensive language generation** (Section 5.1) These comments are examples of non-abusive language. Please generate an example of non-abusive language related to the following Context in Korean like a Korean internet user.
	Korean + Context (Without) 앞의 comment들은 언어폭력의 예시이다. 한국인 인터넷 유저처럼 언어폭력을 한국어로 생성하라. **Korean + Context (With)** 앞의 comment들은 언어폭력의 예시이다. 한국인 인터넷 유저처럼 주어진 Context에 관련된 언어폭력을 한국어로 생성하라. **Korean + Context (With) for not-offensive language generation** (Section 5.1) 앞의 comment들은 비언어폭력의 예시이다. 한국인 인터넷 유저처럼 주어진 Context에 관련된 비언어폭력을 한국어로 생성하라.
Context	**Without**
	With Context: 멘션 스루 자주 있어용..!!! 고의가 아니니 오해하지 말아주세용 (I often go through mentions..!!! It's not intentional, so please don't get me wrong)
Guidance	Comment:

Table 10: The statistical results for automatic scores. For the analysis, we conducted ANOVA test for mean comparison and Tukey's HSD test for Post Hoc test.

Models	Metrics	Conditions	M	SD	F	p	Subset for alpha = 0.05 1	2	3	4
gpt-3.5-turbo	Self-B	$P_{E,E,w/o}$	2.17	0.37	41.430	.000			2.17	
		$P_{E,E,w}$	1.83	0.16				1.83		
		$P_{E,K,w/o}$	2.37	0.98						2.37
		$P_{E,K,w}$	1.61	0.13			1.61			
		$P_{K,E,w/o}$	2.02	0.29					2.02	
		$P_{K,E,w}$	1.76	0.17			1.76	1.76		
		$P_{K,K,w/o}$	1.70	0.31			1.70	1.70		
		$P_{K,K,w}$	1.69	0.19			1.69	1.69		
	Tox.	$P_{E,E,w/o}$	.7218	.1991	70.718	.000	.7218			
		$P_{E,E,w}$	.6837	.2103			.6837	.6837		
		$P_{E,K,w/o}$	.2351	.2369						.2351
		$P_{E,K,w}$	.4143	.2810					.4143	
		$P_{K,E,w/o}$	.7112	.2048			.7112			
		$P_{K,E,w}$	.6000	.2188				.6000		
		$P_{K,K,w/o/}$	.3621	.2115					.3621	
		$P_{K,K,w}$	.3747	.2192					.3747	
text-davinci-003	Self-B	$P_{E,E,w/o}$	2.35	0.53	132.629	.000			2.35	
		$P_{E,E,w}$	1.56	0.21			1.56			
		$P_{E,K,w/o}$	1.55	0.14			1.55			
		$P_{E,K,w}$	1.48	0.20			1.48			
		$P_{K,E,w/o}$	1.72	0.24				1.72		
		$P_{K,E,w}$	1.52	0.20			1.52			
		$P_{K,K,w/o}$	1.46	0.19			1.46			
		$P_{K,K,w}$	1.49	0.13			1.49			
	Tox.	$P_{E,E,w/o}$	.8529	.1709	59.805	.000	.8529			
		$P_{E,E,w}$	.5666	.2827				.5666		
		$P_{E,K,w/o}$	.6091	.2937				.6091		
		$P_{E,K,w}$	.2864	.2937						.2864
		$P_{K,E,w/o}$	.6410	.2776				.6410		
		$P_{K,E,w}$	.4211	.2440					.4211	
		$P_{K,K,w/o}$	.4094	.2430					.4094	
		$P_{K,K,w}$	.2920	.2362						.2920
polyglot-ko-5.8b	Self-B	$P_{E,E,w/o}$	3.38	4.18	10.513	.000		3.38	3.38	
		$P_{E,E,w}$	2.67	4.13				2.67	2.67	
		$P_{E,K,w/o}$	3.87	4.41					3.87	3.87
		$P_{E,K,w}$	5.04	6.94						5.04
		$P_{K,E,w/o}$	2.94	3.19				2.94	2.94	
		$P_{K,E,w}$	1.82	2.60			1.82	1.82		
		$P_{K,K,w/o}$	2.08	2.00			2.08	2.08		
		$P_{K,K,w}$	0.97	0.59			0.97			
	Tox.	$P_{E,E,w/o}$	.2542	.2936	7.916	.000	.2542			
		$P_{E,E,w}$	.1769	.2030			.1769	.1769	.1769	
		$P_{E,K,w/o}$	.1975	.2356			.1975	.1975		
		$P_{E,K,w}$	.1585	.2166				.1585	.1585	.1585
		$P_{K,E,w/o}$	.1221	.1597				.1221	.1221	.1221
		$P_{K,E,w}$	.1163	.1335				.1163	.1163	.1163
		$P_{K,K,w/o}$	.1104	.1417					.1104	.1104
		$P_{K,K,w}$	.0887	.1043						.0887

Table 11: The statistical results for human evaluation scores.

Metrics	Conditions	M	SD	F	p	Subset for alpha=0.05 1	2	3	4
Relevance	$P_{E,E,w/o}$	-		2.341	.078				
	$P_{E,E,w}$	4.400	.8165						
	$P_{E,K,w/o}$	-							
	$P_{E,K,w}$	3.600	1.323						
	$P_{K,E,w/o}$	-							
	$P_{K,E,w}$	3.880	1.236						
	$P_{K,K,w/o}$	-							
	$P_{K,K,w}$	4.167	1.090						
Offensiveness	$P_{E,E,w/o}$	3.880	1.003	10.472	.000	3.880	3.880	3.880	
	$P_{E,E,w}$	4.320	.7407			4.320	4.320		
	$P_{E,K,w/o}$	2.240	1.172						2.240
	$P_{E,K,w}$	3.480	1.460				3.480	3.480	
	$P_{K,E,w/o}$	4.480	.7068			4.480	4.480		
	$P_{K,E,w}$	4.520	.5799			4.520			
	$P_{K,K,w/o}$	3.640	1.336			3.640	3.640	3.640	
	$P_{K,K,w}$	3.240	1.287					3.240	
Fluency	$P_{E,E,w/o}$	3.280	1.230	2.151	.040		3.280		
	$P_{E,E,w}$	4.040	1.195			4.040	4.040		
	$P_{E,K,w/o}$	3.440	1.280			3.440	3.440		
	$P_{E,K,w}$	3.880	1.319			3.880	3.880		
	$P_{K,E,w/o}$	3.640	1.025			3.640	3.640		
	$P_{K,E,w}$	4.080	1.066			4.080	4.080		
	$P_{K,K,w/o}$	3.840	1.017			3.840	3.840		
	$P_{K,K,w}$	4.320	1.168			4.320			